

Linear Regression

ba phân tích chính trong DA

- + Khai báu: (analysis of difference)
- + Lý quan (association analysis)
- + Tương quan (correlation analysis) & tín hiệu (prediction)

- ① phân tích khai báu: t-test, ANOVA, Z-test, Chi-square
- ② phân tích lý quan: odds ratio, Risk ratio, Prevalence ratio
- ③ phân tích tương quan & tín hiệu:

- Correlation analysis
- Linear Regression
- Logistic Regression.
- Cox's Regression

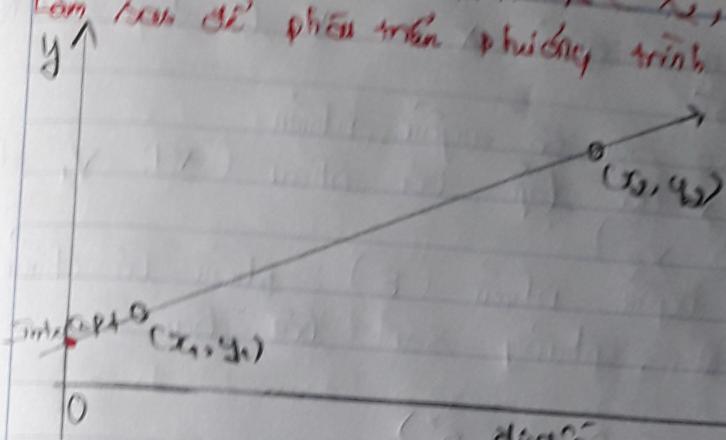
Phân tích tương quan:

- Danh giá mối tương quan
- Tỉ số tương quan (Coefficient of correlation) viết tắt là r
- Chứng ta cần biết thêm
 - Mức độ ảnh hưởng của biến tín hiệu (predictor variable) trên biến phụ thuộc (dependent variable)
 - Tín hiệu

Mục tiêu Linear Regression model.

- Gìn một mô hình (phép trình) mô tả mối liên quan giữa $x \rightarrow y$
- x có thể là độ tuổi, trọng lượng.
- y có thể là BMD (mật độ xương)
- Đưa chính xác yếu tố nhiều
- Tín hiệu

* Cho 2 điểm (x_1, y_1) & (x_2, y_2)
Làm sao để phác thảo phương trình rời rạc này?



① Góc Gradient (Slope) $\text{Slope} = \frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1}$

② Góc giá trị khởi đầu (Intercept) (của y khi x=0)

Nó chính là quay trung bình:

- y là biến phụ thuộc (response variable, dependent variable)
 - y phải là biến liên tục.
- x là biến độc lập (predictor predictor variable, independent variable ...)
 - x là biến liên tục hoặc không liên tục
- Nó chính!

$$y = \alpha + \beta x + \epsilon$$

α : intercept.

β : Slope / gradient

ϵ : Sai số ngoài nhanh (Random error - Nguồn
đao động của y trong mỗi giá trị của x)

Giai định:

- Nối liên quan giữa X và Y là tuyến tính (linear)
- vẽ tham số
- X không có sai số ngẫu nhiên
- Giai thị y độc lập với nhau (Vd: y_1 không liên quan với y_2)
- Sai số ngẫu nhiên (ε): phân bố chuẩn, trung bình 0, phương sai bất biến

$$\varepsilon \sim N(0, \sigma^2)$$

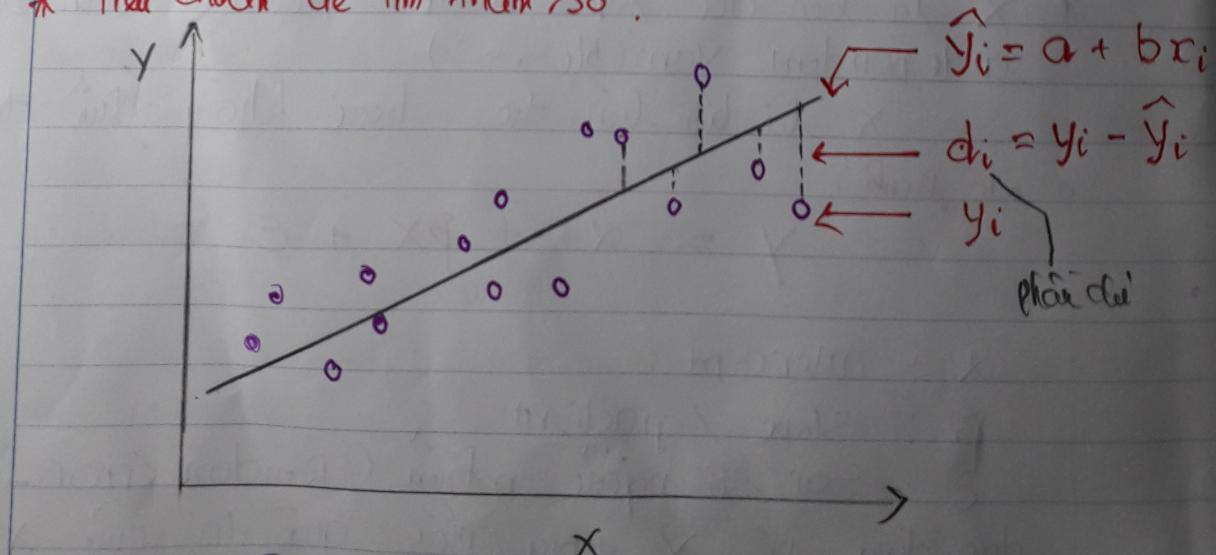
Ước tính tham số (Parameters)

Mục tiêu:

o Mô hình: $y = \alpha + \beta x + \varepsilon$

- Chúng ta không biết α & β
- Nhưng có thể dùng dữ liệu thực nghiệm để ước tính α & β để
- Ước số (estimate.) của α & β là a và b.

* Trảm chuẩn để tìm tham số:



Gim Công thức (Estimator) để tính a và b sao cho $\sum d_i^2$ là nhỏ nhất \rightarrow Least square - Bằng phẳngely nhés nhést

vô tính bằng R.

Thứ
Ngày

No.

Chúng ta muốn xác định mối liên quan BMD và trọng lượng.

Mô hình Rồi quy tuyến tính.

$$BMD = \alpha + B \text{ weight} + \epsilon$$

R là lm ($BMD \sim \text{weight}$)

Điều gì kết quả:

Coefficients	Estimate	Std. Error	t Value
(Intercept)	0.381	0.0138	27,55
wt	0.063	0.0001	32,89

Nó hình là:

$$BMD = a + b * \text{weight}$$

Phương trình

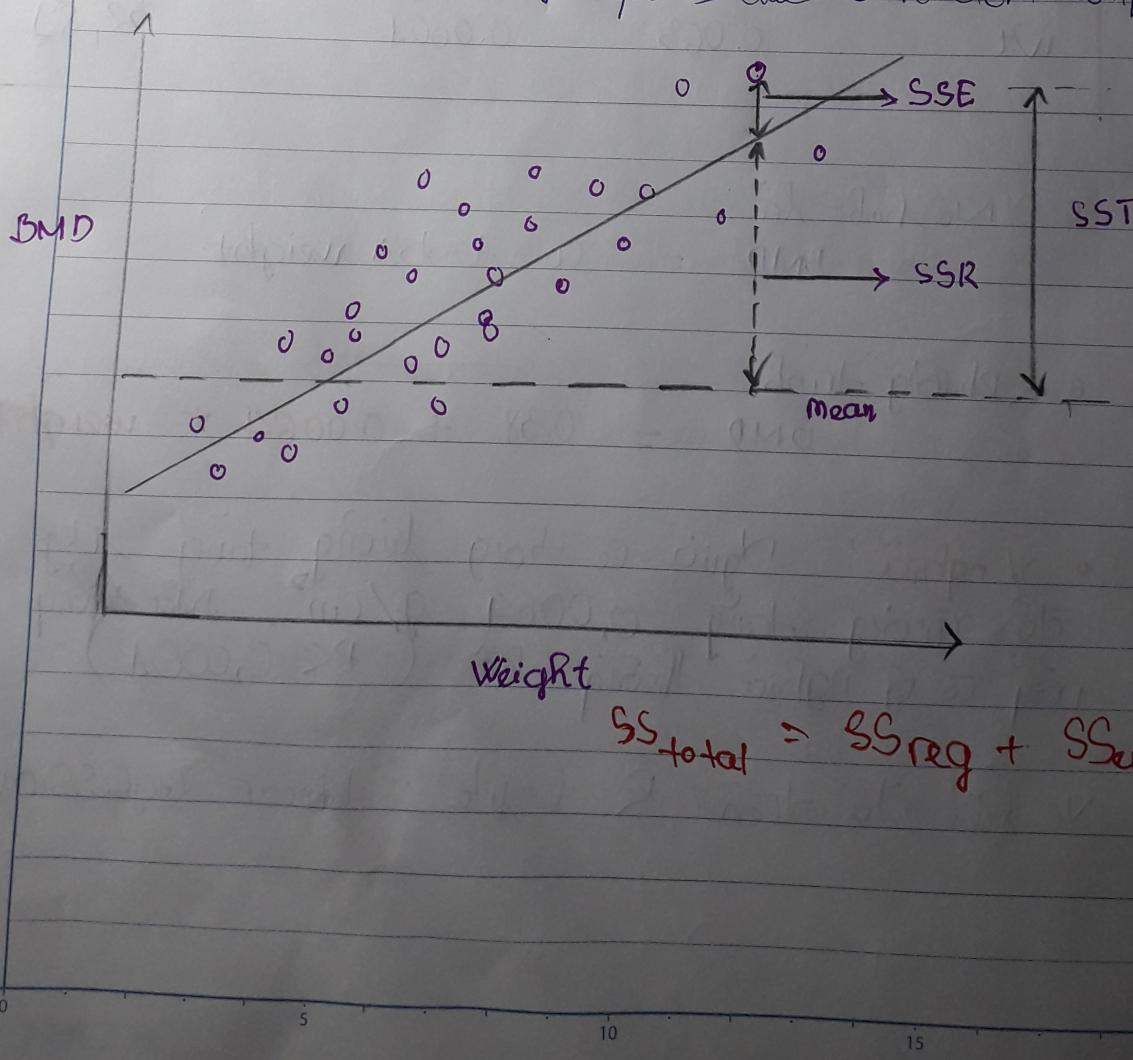
$$BMD = 0.38 + 0.0064 * \text{weight}$$

Ý nghĩa: Người có trọng lượng tăng 1kg thì mật độ xương tăng 0,0064 g/cm². Nói通俗 quan trọng này có ý nghĩa thống kê ($p < 0,0001$)

Vẽ biểu đồ trên R. (file: Linear Regression)

Mô hình hồi quy tuyến tính: Phân tích phương sai.
Phân tích phương sai:

- $BMD = a + b * weight + e$
- $BMD = a + b * weight + e$
- Observed Variation = model + Random
"Variation" = Sum Of Squares.
- $SS_{total} = \text{total sum of squares}$
- $SS_{reg} = \text{sum of squares due to the regression model}$
- $SS_{error} = \text{sum of squares due to random component}$



Bảng (*)

R^2

Phân tích biến phân佈 方差分析

Thứ
Ngày

Ngày

R \Rightarrow Analysis of variance table.

Bảng (*) Analysis of variance table

Response : gnbmid

	Df	Sum Sq	Mean Sq	F Value	Pr(>F)
Wt	1	17,153	17,1528	1081,7	$< 2,2 \times 10^{-16}$
Residuals	2100	33,616	0,0159		

o Total SS = 17,15 + 33,62 = 50,77

- Do weight : 17,15

+ Phản ứng trừu tượng (residuals) : 33,62.

Hệ số xác định R^2

$$R^2 = \frac{17,15}{50,77} = 0,34$$

\rightarrow Biến cần năng (Wt) giải thích 34% nhưng khác biệt về mức độ xứng 1/1 Gé Cát nhân.

Hệ số xác định điều chỉnh (adjusted R^2)

o Định nghĩa dễ hiểu nhất.

$$R^2_{adj} = 1 - \left(\frac{MS_{error}}{MS_{total}} \right)$$

MS_{error} : mean square due to error.

MS_{total} : mean square total

theo bảng (*)

o $MS_{total} = (17,15 + 33,62) / 2121 = 0,0239$

o $MS_{error} = 0,0159$

o $R^2_{adj} = 1 - (0,0159 / 0,0239) = 0,337$

Phương Sai Cuối BMD Sau Khi Chia Chính ict.
Bảng (*)

o Mean Square (MS) = Sum of Square / degrees
of freedom

o MS (Residuals) = $33,616 / 120 = 0,0159$.

Phương Sai Sau Khi Chia Chính
trong là 0,0159. (trước Khi Chia Chính: 0,0239)

Tên Lượng Tốt hơn:

- Nếu hình là: $BMD = 0.38 + 0.0064 * weight$
- Nếu không bao trong (weight), BMD trung bình
là 0.83 g/cm^2 .
- Nếu bao trong lượng, chia về tien lượng tốt hơn
- $weight = 50 \text{ kg}$, $BMD = 0.38 + 0.0064 \cdot 50$

$$= 0.70 \text{ g/cm}^2$$

Mô hình hồi quy tuyến tính: Phân tích dãy đồng dư
(Residual analysis)

Residuals - dãy đồng dư:
 $y = \hat{a} + \hat{b}x + e$

Gia trị trung bình $E(y) = \hat{y} = \hat{a} + \hat{b}x$
phản dư: $e = y - \hat{y}$

Đữ liệu quan sát = mô hình tiên lượng + phản dư (nhầm)
phản dư = Gia trị quan sát - Gia trị tiên lượng

Mục đích phân tích dãy đồng dư:

- Kiểm định phản bội chuẩn (normal distribution)
- Phân tích sự cố phản bội biến vs x ?
- Độ lặp?
- Có giá trị nào là "ngẫu vi" (outlier) hay có ảnh hưởng (Influential observation)

Dãy đồng dư chuẩn hóa:

- Dãy đồng dư tuy thuộc vào dạng vi phân
- Mỗi giá trị của dãy đồng dư là $\bar{x} + phuong sai$ và là lệch chuẩn
- Chuan Hoa = standardization.

$$\text{Standardized Residual } i = \frac{\text{Residual } i}{\text{Standard Deviation of Residual } i}$$

hàm $x_{\text{cố định}} R' \cdot \text{standard}(C)$

Kiểm tra phương sai: (Homogeneity)

Thứ
Ngày

No.

- Giả định 2: Phương sai của mỗi giá trị trên lượng trưng đường nhau (Homogeneity)
- Phương pháp kiểm tra
 - NCV Test
 - Biến đổi Studentized residual & giá trị biến

(R)

$$m_1 = lm(fnbmd ~ wt)$$

library(car)

NCVTest(m1)

Studentized residuals * giá trị biến lũy
Spread Level Plot (m1)

kiểm tra tuyết tính:

- Giả định: mỗi biến quan phai tuân thủ (Linear)
- Phương pháp kiểm tra: partial residual plot
(còn gọi là Component + residual plot)
- Kiểm tra độc lập: Independence
- (R): durbinWatsonTest(m1)

Outliers

- (R) OutlierTest(m1)
- qq Plot (m1)
- leverage Plot (m1)

Kiểm tra giá trị ảnh hưởng: Influential observations.

(R)

Thứ
Ngày

No.

CoolIC's D plot

identify D values $> 4/(n-k-1)$

$\rightarrow \text{Cutoff} = 4 / ((\text{nrav(dat)} - \text{length(m1\$coefficients)}) - 2)$

Mô hình hồi quy tuyến tính đa biến.

Nội dung:

- Biến phân nhánh (Category predictors)
- Nhàm hồi quy đơn giản vs biến phân nhánh
- Mô hình hồi quy với 2 biến تكون lượng

Biến phân nhánh: (categorical variables)

Nominal: tinh danh

Ordinal: Thứ tự

- Biến tinh danh: giới tính, chủng tộc, địa điểm.
- Thứ tự: giới đoạn kinh (I, II, III...)

Hồi quy tuyến tính như là một t-test

Chúng ta đã biết kiểm định dùng t để phân tích so sánh mứa biến liên tục giữa 2 nhóm

(R) t-test (age ~ group)

- nhưng chúng ta cũng có thể so sánh 2 biến phân

tích hồi quy tuyến tính.

$$y = \alpha + \beta x + \epsilon$$

Tương ứng với x là biến phân nhánh

• VD: mật độ xương (BMD) l' Nam & nữ

Gọi: giới tính là biến Gender, Chênh lệch là

$$BMD = \alpha + \beta(\text{gender}) + \epsilon$$

(R)

$$\text{lm}(\text{fn.BMD} \sim \text{gender})$$

Điều kiện kết quả:

→ Summary ($\text{lm}(\text{fn.BMD} \sim \text{gender})$)

Coefficients:

(Intercept) Estimate
0,727463

Gender_Male Estimate
0,13195

R-Square: 0,1729

* Hết số hồi quy (Regression Coefficient) = 0,13, có ý nghĩa là Male (α) BMD cao hơn nữ giòn 0,13 g/cm².
* Số cách khác, 0,13 chính là chênh lệch BMD giòn nam -> nữ ($0,91 - 0,78 \text{ g/cm}^2$)

* Hết số xác định (R-Square): 0,17, công thức biến gender giải thích 17% dao động về BMD l' l' cao cá nhân.

Nên hình hồi quy tuyến tính đa biến.

(R)

$$BMD = \alpha + \beta_1 * \text{gender} + \beta_2 * \text{weight}$$

→ m = lm(fn.BMD ~ gender + wt)

kết quả:

Coefficients:

Category Variable

	Estimate	Std. Error	t-value
(Intercept)	0,4251819		
Gender_Male	0,0608661		
wt	0,00541218		

$$BMD = 0,425 + 0,061 * \text{Male} + 0,0054 * \text{Weight}$$

Nữ vs nam:

$$\begin{aligned} BMD &= 0,425 + 0,061 * \text{Male} + 0,0054 * \text{Weight} \\ &= 0,486 + 0,0054 * \text{Weight} \end{aligned}$$

Đối vs nữ:

$$BMD = 0,425 + 0,0054 * \text{Weight}$$

Điều chỉnh:

* Sau khi điều chỉnh cho trong luồng, BMD ở Nam cao hơn nữ 0,061 g/cm² ($P < 0,0001$)

* Chiều cao: Trước khi điều chỉnh, mức độ khác biệt là 0,13 g/cm².

MLR 2:

Vết Tính

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_p x_p + \epsilon$$

Y là biến phụ thuộc (dependent variable), biến liên tục
 $x_1, x_2, x_3, \dots, x_p$: Biến tiên lượng

$b_1, b_2, b_3, \dots, b_p$: Regression Coefficients (hệ số hồi quy)

Tham số:

Nhàm (***)

Dùng để tìm thức để xác định B ở mô hình (***)

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_p x_p$$

- $b_1, b_2, b_3, \dots, b_p$ là các số $\hat{b}_1, \hat{b}_2, \hat{b}_3, \dots, \hat{b}_p$

$$\epsilon = \hat{y} - \bar{y} = (\text{Residuals})$$

Phương pháp least squares

Least squares method - phương pháp bình phương sai thiểu

• Hàm least square

$$L = \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \sum_{j=1}^k \beta_j x_{ij})^2$$

• Ứng số least square phỏa đập ứng.
Tmf đối hàn.

Nhân ngữ ma + rãnh:

$$y = X \cdot \beta + \epsilon$$

Giai đt:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}; X = \begin{bmatrix} 1 & x_{11} & \dots & x_{1k} \\ 1 & x_{21} & \dots & x_{2k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \text{ and } \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\text{Đáp số: } \hat{\beta} = (X'X)^{-1}X'y$$

Hàm lm trong R.

• Trong R, có hàm lm (linear models)

- Giải phương trình để tìm tinh tham số

Giải toán bài chí sít thông kê lùn quan đến mô hình

• Định giá sai thul hòn của mô hình

* Công thức Cherny

$$\text{lm}(Y) \approx x_1 + x_2 + x_3 + \dots$$